Kingman's Unlabeled n-Coalescent

Raazesh Sainudiin[○]

[○] **Biomathematics Research Centre**

Department of Mathematics & Statistics, University of Canterbury, NZ,

[‡] Department of Statistics, University of Oxford, UK, and

[⊛] Biological Sciences, University of California, Irvine, USA.

Joint with: Peter Donnelly[‡], Bob Griffiths[‡], Gil McVean[‡], and Kevin Thornton[⊛]

Outline – Talk Outline

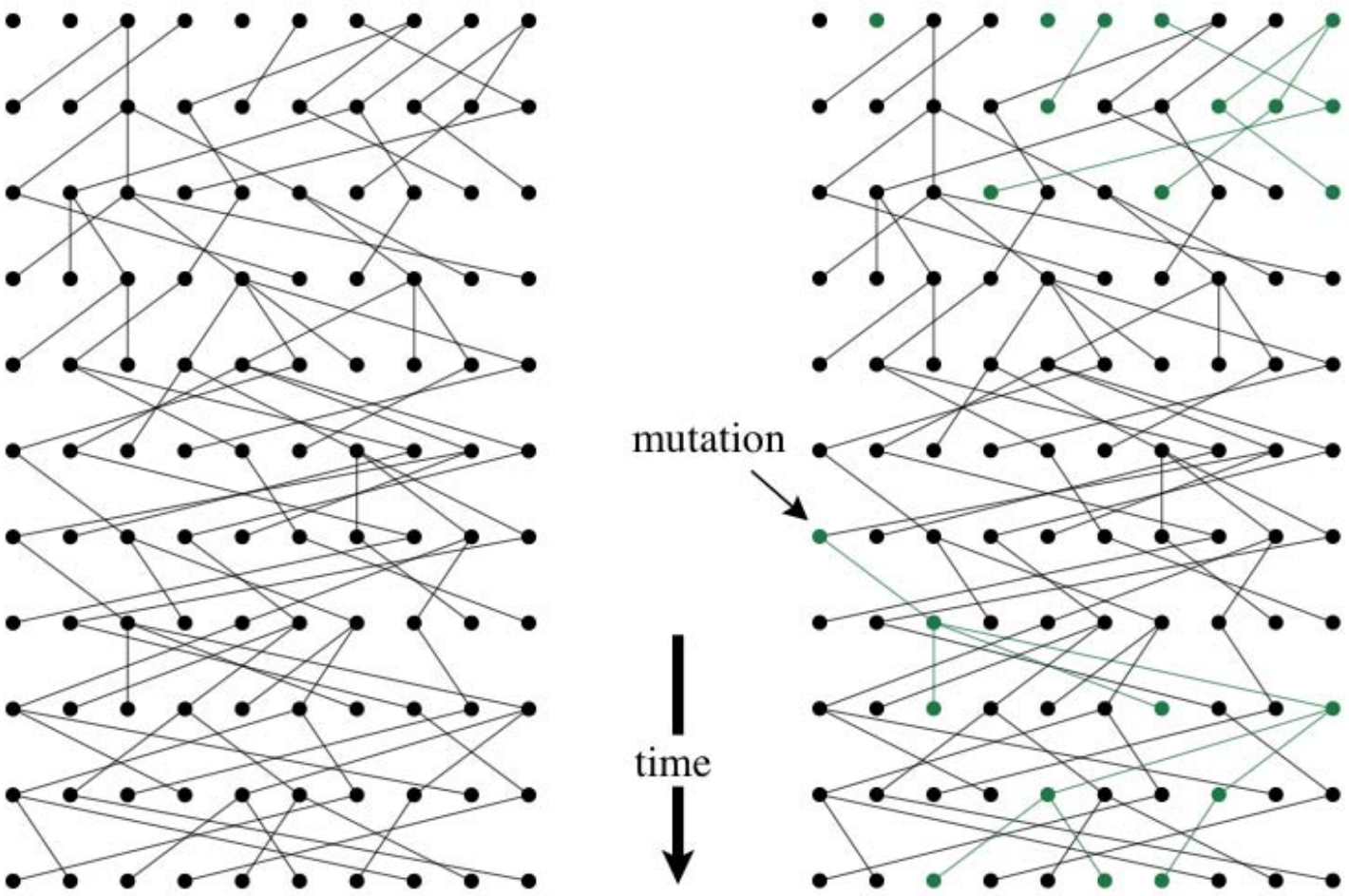
- The Coalescent Models
- Computationally Intensive Likelihoods
- A Paritllay-ordered Coalescent Experiments Graph
- Unlabeled n-Coalescent
- Results
- Summary
- Acknowledgments

Data and Model 1: $\phi \equiv \theta \in \Phi, \theta = 4N_e\mu$ (scaled mutation rate)

The Wright-Fisher Model – Random Mating, Constant Size, No Recombination/Selection

A Population of $N = 10$ homologous DNA seqns. of length m and the Population History of site i

```
  : 1 2 3 4 5 6 7 8 9 10
1 : A A A A A A A A A C
2 : G G G G G G G G G G
...
i : T T A A A A A A A A
...
k : ...
```

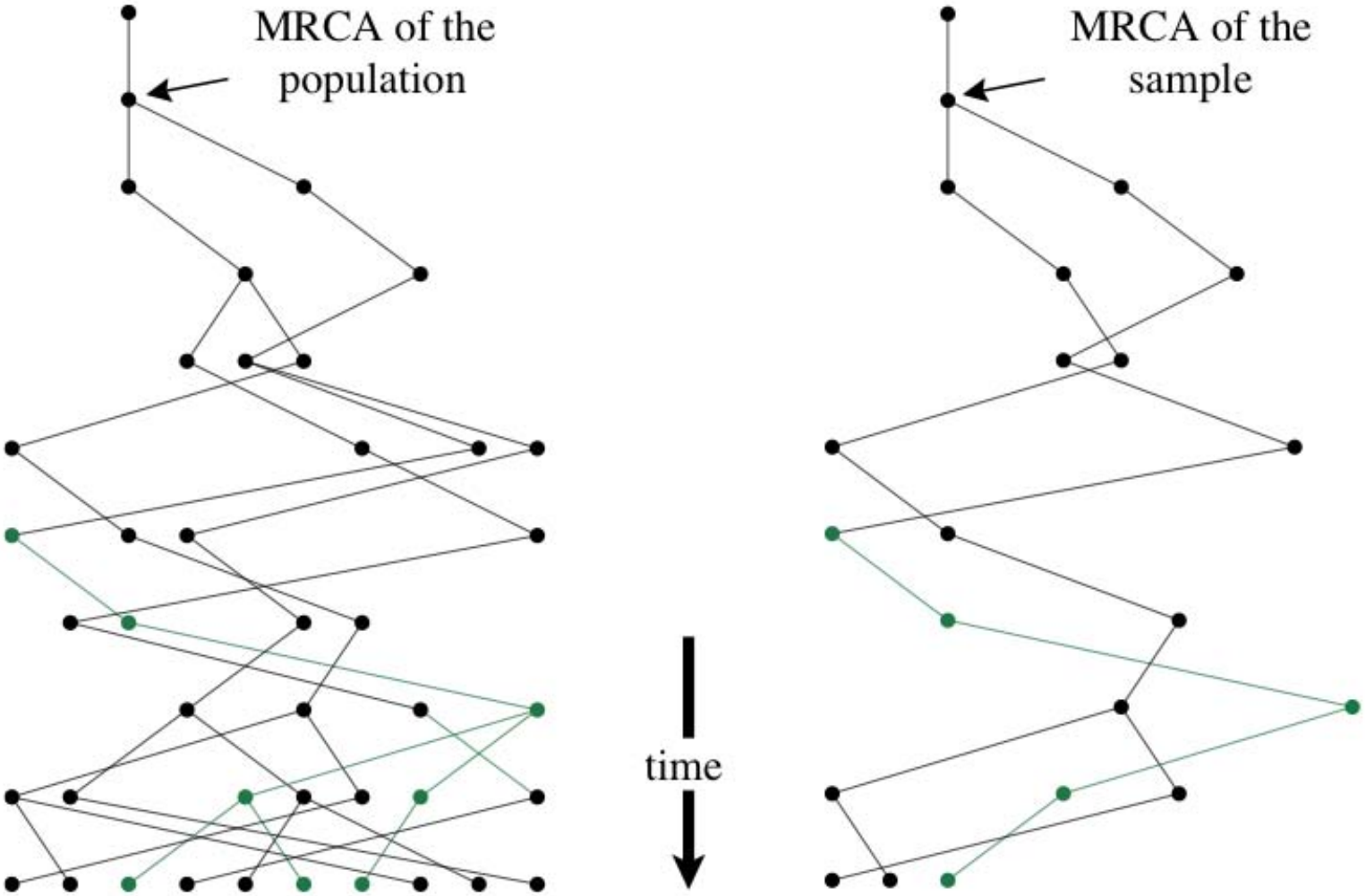


Data and Model 1: $\phi \equiv \theta \in \Phi$, $\theta = 4N_e\mu$ (scaled mutation rate)

The Wright-Fisher Model – Random Mating, Constant Size, No Recombination/Selection

Ex: Data of 3 homologous DNA sequences at site i, its Population History and the Sample History of sampled individuals 1,2, and 3.

```
  : 1 2 3
i : T T A
```



Model 1: $\phi \equiv \theta \in \Phi, \theta = 4N_e\mu$ (scaled mutation rate)

The Coalescent Approximation of the Wright-Fisher (W-F) Model (Kingman, 1982)

A Sample Coalescent Sequence or c-sequence ($\{\{1\},\{2\},\{3\}\}, \{\{1,2\},\{3\}\}, \{\{1,2,3\}\}$)

and coalescent times or epoch times $t_i, i \in \{3,2\}$.

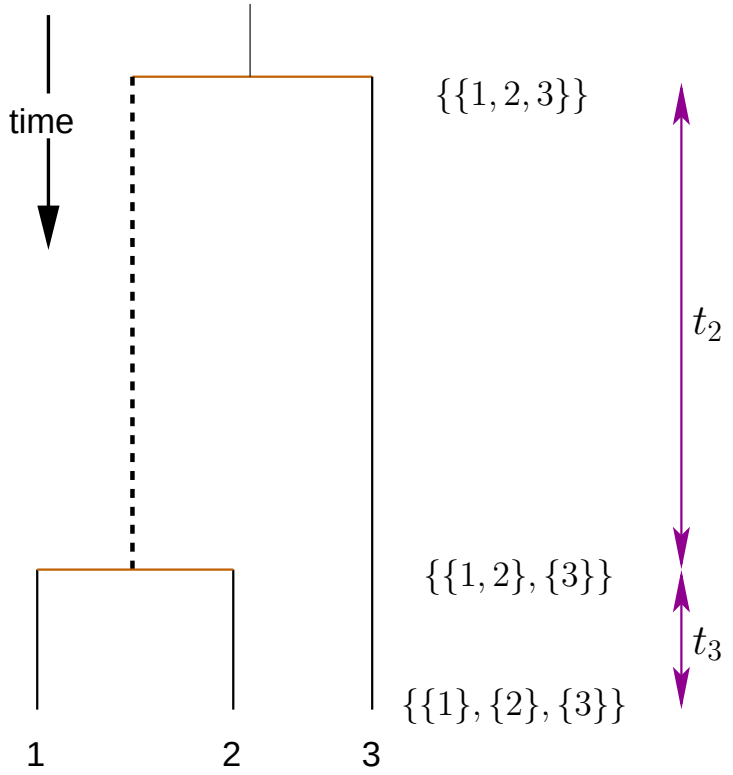
- Offspring "choose" parents uniformly and independently in W-F model
- $\Pr(2$ lineages coalesce in 1 generation) = $1/N$
- $\Pr(2$ lins. are distinct $> g$ gens.) = $(1-1/N)^g$
- Rescaled time t is g in units of N gens. Then, $\Pr(2$ lins. remain distinct $> t$) is

$$(1-1/N)^{\lfloor Nt \rfloor} \overset{N\to\infty}{\longrightarrow} e^{-t}$$

- Lineage Death Process: In general, the R.V. T_i that any pair of i lineages coalesce is approximately exponentially distributed for large N.

$$T_i \sim \text{Exponential}\left(\binom{i}{2}\right)$$

- Uniform Binary Fusion of two extant lineages.



Model 1: $\phi \equiv \theta \in \Phi, \theta = 4N_e\mu$ (scaled mutation rate)

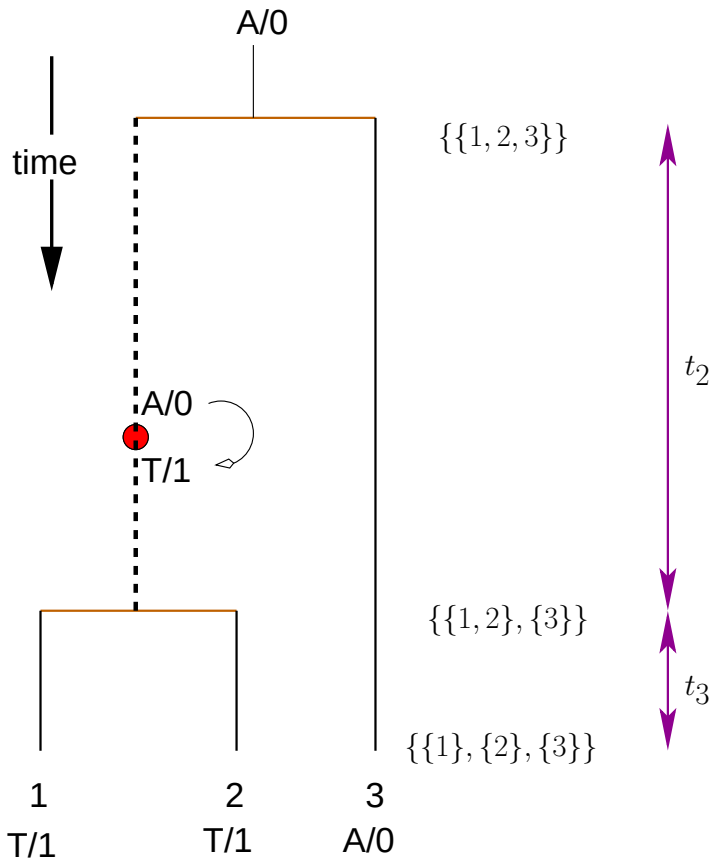
The Coalescent Approximation of the Wright-Fisher (W-F) Model (Kingman, 1982)

- The n-Coalescent is a continuous time Markov Chain on $\mathbb{C}_n \equiv \cup_{i=1}^{n} \mathbb{C}_n^i$, the set partitions of $\{1, \ldots, n\}$, with rates $q(c_h|c_g),\ c_g, c_h \in \mathbb{C}_n$:

$$q(c_h|c_g) = \begin{cases} -i(i-1)/2 & : \text{if } c_g = c_h \in \mathbb{C}_n^i \\ 1 & : \text{if } c_h \succ_c c_g \\ 0 & : \text{o.w.} \end{cases}$$

$c_h \prec_c c_g \Leftrightarrow c_h = c_g \setminus c_{g,j} \setminus c_{g,k} \cup (c_{g,j} \cup c_{g,k})$

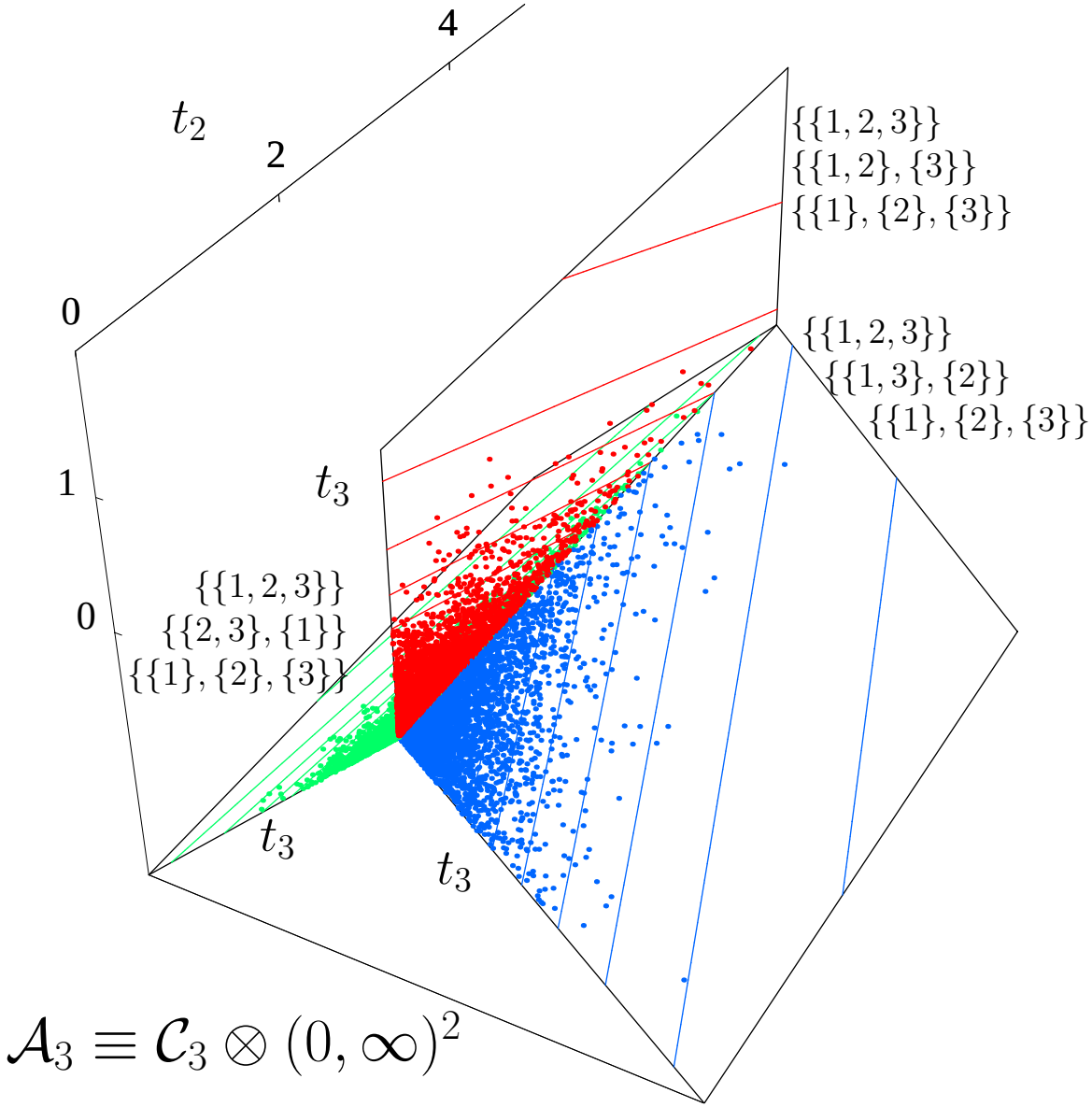
a realization $c = (c_n, c_{n-1}, \ldots, c_1) \in \mathcal{C}_n$



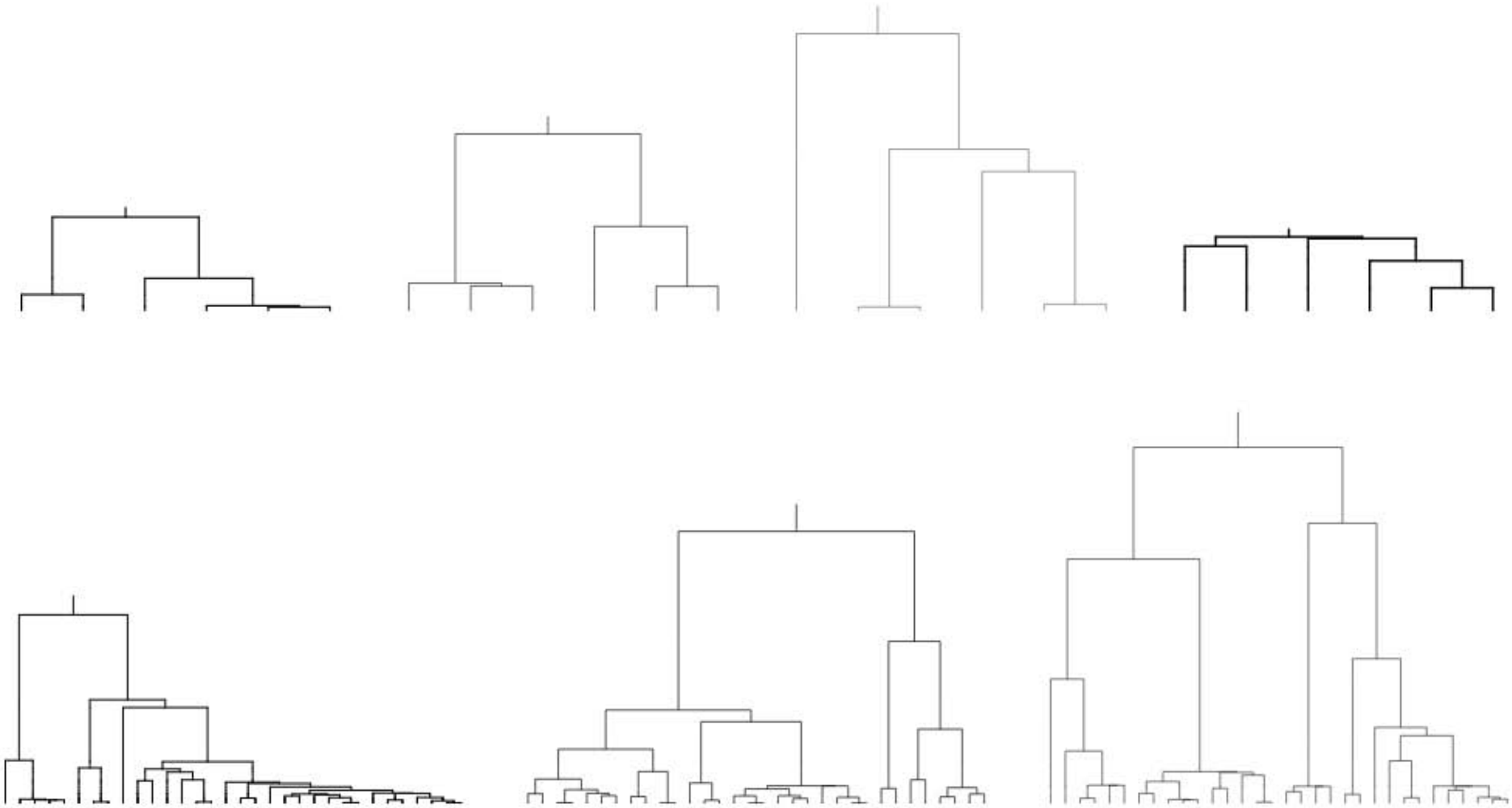
- Superimpose indep. mutations $\sim Poisson(\theta/2 \equiv 2N\mu)$

∞-many-sites mutation model

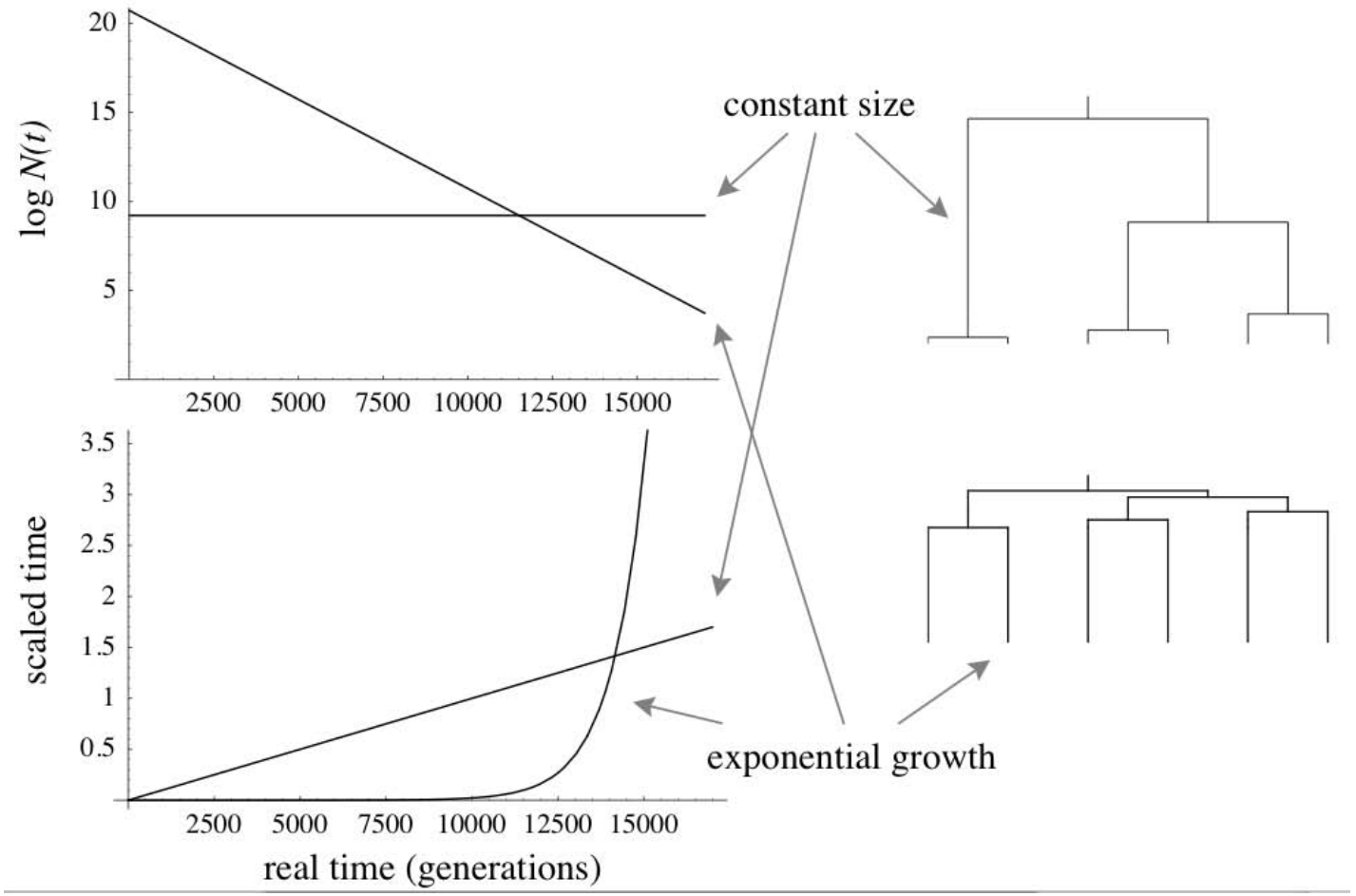
Coalescent Space $\mathcal{A}_n \equiv \mathcal{C}_n \otimes (0,\infty)^{n-1}$ when $n = 3$ (Model 1)



Realizations from $\mathcal{A}_n \equiv \mathcal{C}_n \otimes (0,\infty)^{n-1}$ under Model 1, $n = 6, 32$

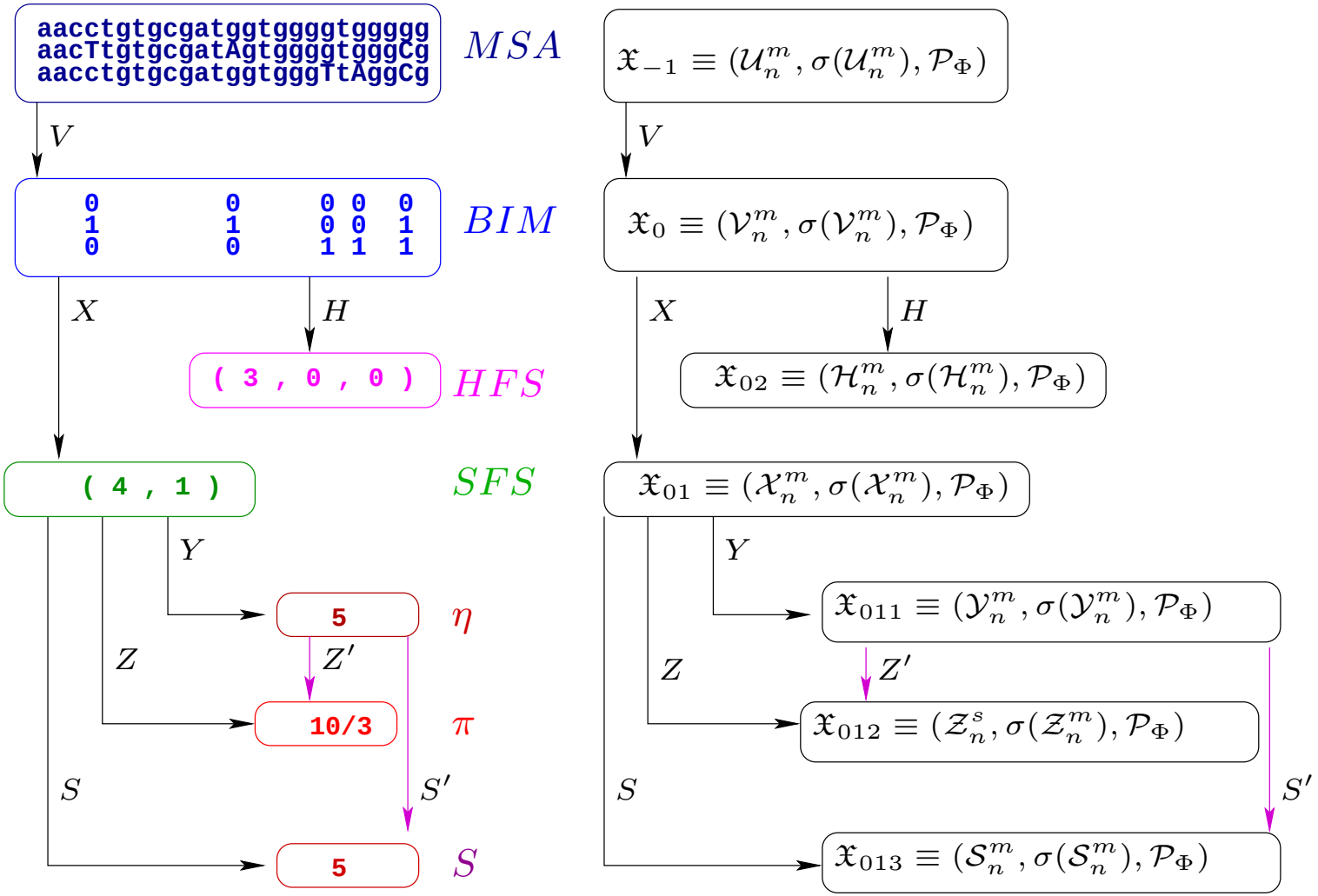


Model 2 : $\phi \equiv (\theta, \nu) \in \Phi$, $\theta = 4N_e\mu$ (scaled mutn. rate) , ν (exp. growth rate)



Figures 1-6 of M. Nordburg, Coalescent Theory, 2000

Coalescent Sample Spaces – Partially Ordered Experiments Graph



(1) Every directed acyclic subgraph of the POEG indexes a Martingale

(2) Each node of the POEG is a tri-sequential asymptotic family of Experiments

Likelihood

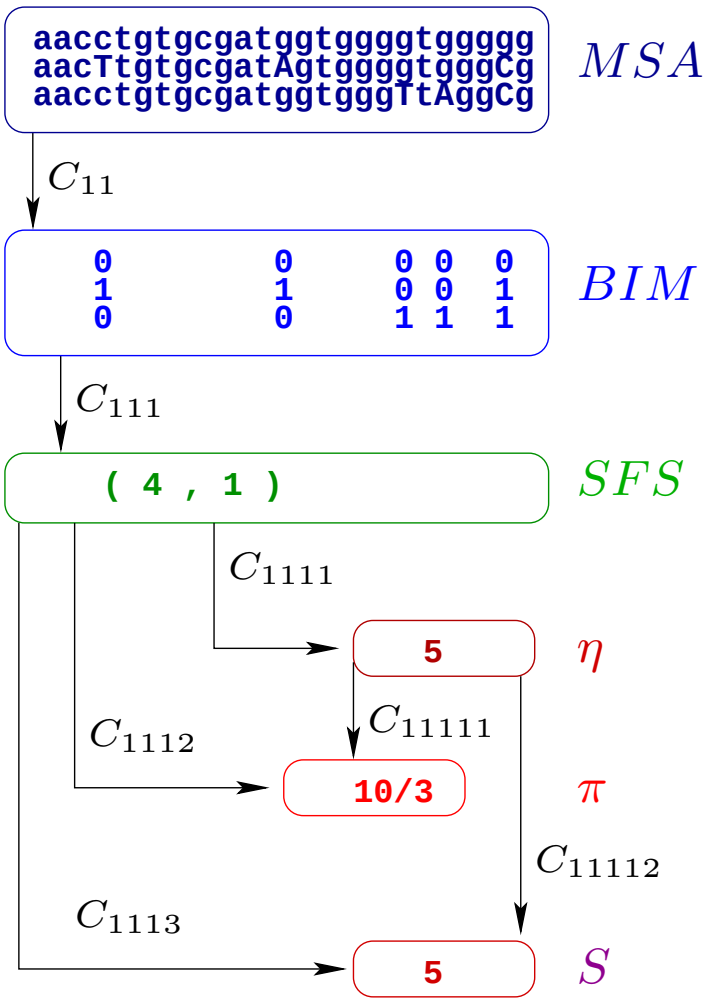
Likelihood, $P(D|\phi)$, is computed by Integrating Missing-Data:

$$\sum_{c \in \mathcal{C}_n} \int_{t \in (0,\infty)^{n-1}} P(D|c,t,\phi) P(c,t|\phi) dt \; dc$$

Cardinalities of the state spaces of the standard n-coalescent on $\mathbb{C}_n$ and the unlabeled n-coalescent on $\mathbb{F}_n$ (to be seen in the sequel).

n	4	10	30	60	90
$\lvert\mathbb{C}_n\rvert$	15	1.2×10^5	8.5×10^{23}	9.8×10^{59}	1.4×10^{101}
$\lvert\mathbb{F}_n\rvert$	5	42	5.6×10^3	9.7×10^5	5.7×10^7
$\lvert\mathbb{F}_n\rvert/\lvert\mathbb{C}_n\rvert$	0.33	3.6×10^{-4}	6.6×10^{-21}	9.9×10^{-55}	4.0×10^{-94}

Likelihood is computationally prohibitive at MSA/BIM Resolns.



- Exact Methods :

  MSA $10,000$ Auto-validating i.i.d. Posterior Samples in MRS SY2006 – novel
  (3/4 leaved phylogenetic tree spaces)
  ≈ 200 CPU sec for $n <= 3$,
  :-($\rightarrow$ impractical for $n > 4$

  BIM Complete Recursion in PTREE G1980
  (1 Locus, $\theta = 10$, C-Model 1)
  :-($\rightarrow$ out of stack for $n > 4$

- Approximate Methods :

  MSA MCMC in COALESCE KYF1998 : $n < 200$ & heuristic

  BIM SIS in GENETREE GT1994 : $L(\theta|v) \approxeq 4$ CPU hrs / θ

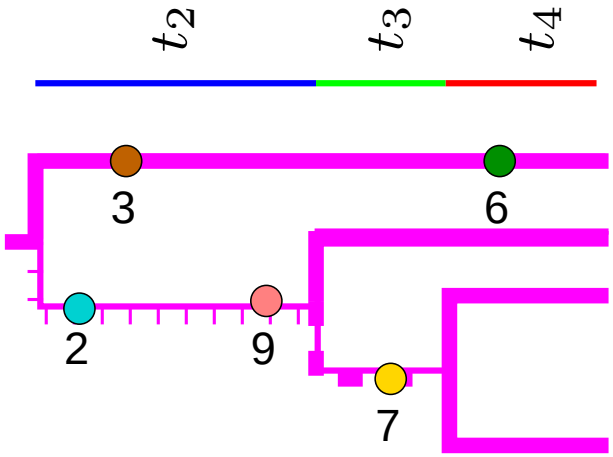
- The Bottom Line: Exact Genome Scanning at fine DNA resolution is currently impractical for $n > 4$

- A Solution: Inference at coarser empirical resolutions, eg. SFS and its sub-experiments – novel

∞-many-sites M-Model: BIM $v \in \mathcal{V}_n^m \to$ SFS $x \in \mathcal{X}_n^m$

Let $v \in \mathcal{V}_n^m \equiv \{0,1\}^{n \times m}$ be a BIM, then the SFS
$x \equiv (x_1, \ldots, x_{n-1}) \in \mathcal{X}_n^m \equiv \{x \in \mathbb{Z}_+^{n-1} : \sum_{i=1}^{n-1} x_i \leq m\}$

$$x_i = N_i(v^T \cdot (1,1,\ldots,1)), \qquad N_i(y_1, y_2, \ldots, y_s) = \sum_{j=1}^{s} \mathbf{1}_{\{i\}}(y_j), \qquad i = 1, \ldots, n-1.$$

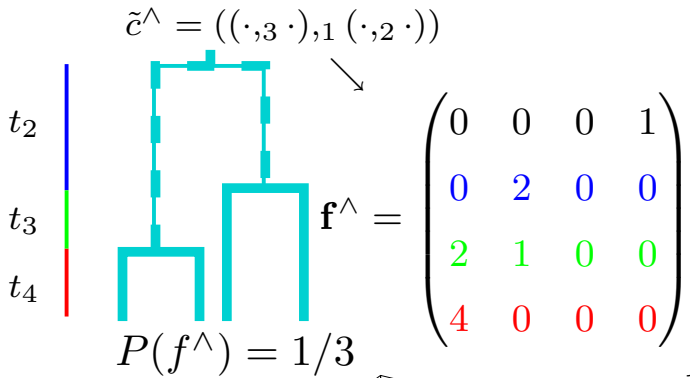


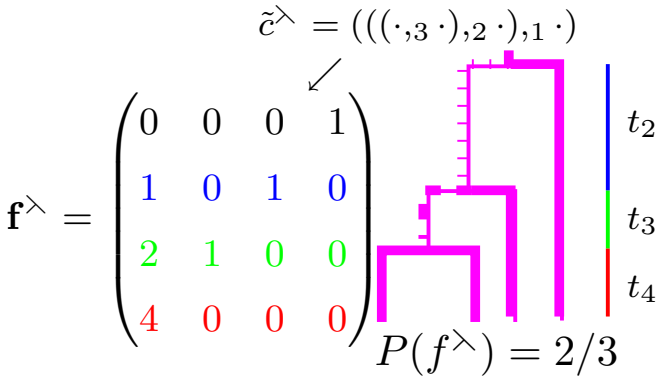
BIM $v \in \mathcal{V}_4^9$

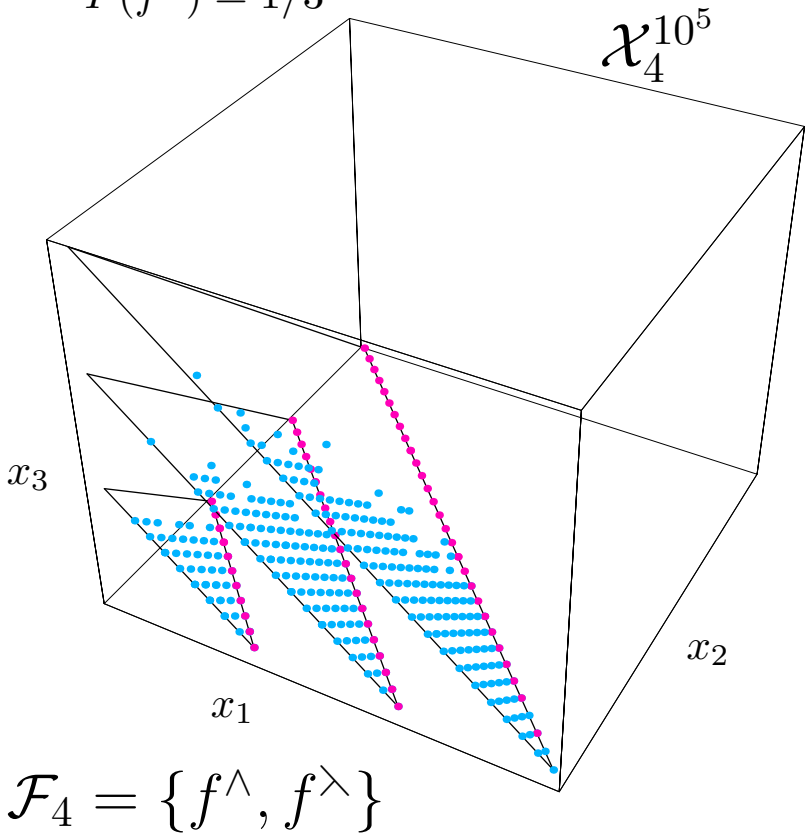
SITES:	1	2	3	4	5	6	7	8	9
IND 1:	0	0	1	0	0	1	0	0	0
IND 2:	0	1	0	0	0	0	0	0	1
IND 3:	0	1	0	0	0	0	1	0	1
IND 4:	0	1	0	0	0	0	1	0	1
COL +:	0	3	1	0	0	1	2	0	3

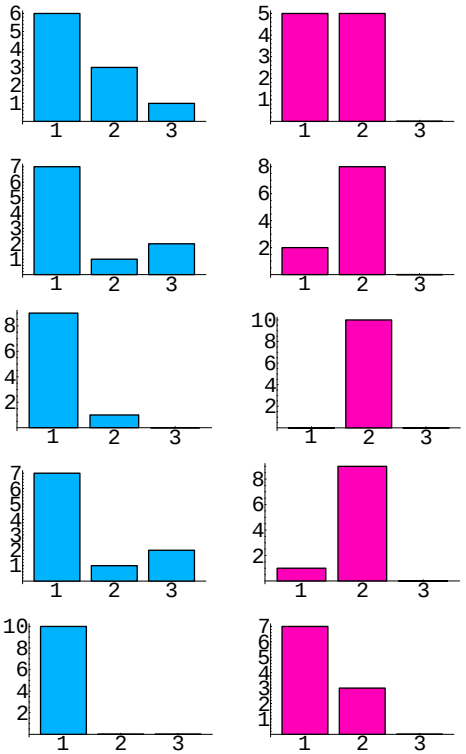
SFS $x = (x_1, x_2, x_3) = (2, 1, 2) \in \mathcal{X}_4^9$

Coalescent Tree Shape, f-Sequence and Site Frequency Spectrum



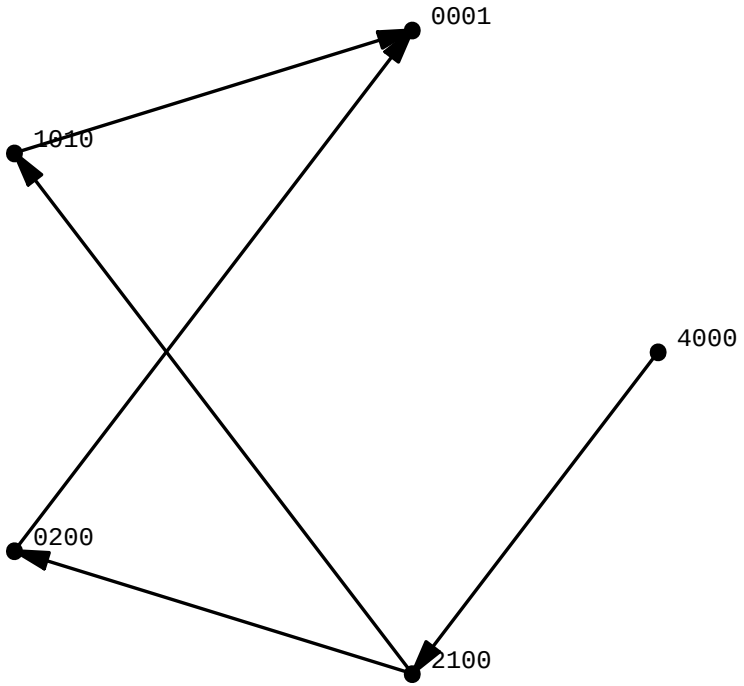




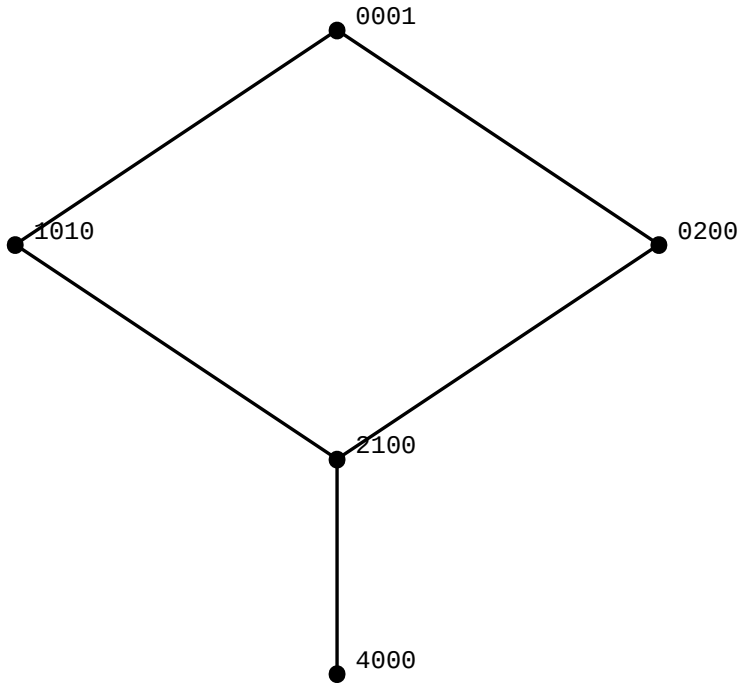


Examples of c-sequence $\rightarrow$ f-sequence, when $n = 4$

Transition-Diagram



Hasse-Diagram



Ex 1:

$[\{1\},\{2\},\{3\},\{4\}],\ [\{1,2\},\{3\},\{4\}],\ [\{1,2,3\},\{4\}],\ [\{1,2,3,4\}] \rightarrow$
$[(4,0,0,0),\ (2,1,0,0),\ (1,0,1,0),\ (0,0,0,1)]$

Ex 2:

$[\{1\},\{2\},\{3\},\{4\}],\ [\{1,2\},\{3\},\{4\}],\ [\{1,2\},\{3,4\}],\ [\{1,2,3,4\}] \rightarrow$
$[(4,0,0,0),\ (2,1,0,0),\ (0,2,0,0),\ (0,0,0,1)]$

Kingman's Unlabeled n-Coalescent

Consider, the integer partitions of n with i blocks:

$$\mathbb{F}_n^i \equiv \{f_i \equiv (f_{i,1}, f_{i,2}, \ldots, f_{i,n}) \in \mathbb{Z}_+^n : \sum_{j=1}^{n} j f_{i,j} = n, \sum_{j=1}^{n} f_{i,j} = i\}.$$

where $f_{i,j}$ denotes the number of lineages subtending j leaves at the i-th epoch.

Proposition (Kingman's Unlabeled n-coalescent)**.** *It is the continuous time Markov chain on $\mathbb{F}_n \equiv \cup_{i=1}^{n} \mathbb{F}_n^i$, the set of integer partitions of n, whose infinitesimal generator $\mathbf{q}(f_h|f_g)$ for any two states $f_g, f_h \in \mathbb{F}_n$ is:*

$$\mathbf{q}(f_h|f_g) = \begin{cases} -i(i-1)/2 & : \text{if} \quad f_g = f_h,\ f_g \in \mathbb{F}_n^i \\ f_{g,j} f_{g,k} & : \text{if} \quad f_h = f_g - e_j - e_k + e_{j+k},\ j \neq k,\ f_g \in \mathbb{F}_n^i,\ f_h \in \mathbb{F}_n^{i-1} \\ (f_{g,j})(f_{g,j}-1)/2 & : \text{if} \quad f_h = f_g - e_j - e_k + e_{j+k},\ j = k,\ f_g \in \mathbb{F}_n^i,\ f_h \in \mathbb{F}_n^{i-1} \\ 0 & : \text{otherwise} \end{cases}$$

Initial state: $f_n = (n, 0, 0, \ldots, 0)$ *and absorbing state:* $f_1 = (0, 0, \ldots, 1)$.

Any realization of the chain is an f-sequence: $f = (f_n, f_{n-1}, \ldots, f_1) \in \mathcal{F}_n$.

Kingman's Unlabeled n-Coalescent

Proposition (Probability of an f_i)**.** *The probability of an $f_i \in \mathbb{F}_n^i$ is:*

$$P(f_i) = \frac{i!}{\prod_{j=1}^{i} f_{i,j}!} \binom{n-1}{i-1}^{-1}$$

Kingman's Unlabeled n-Coalescent

Proposition (Probability of an f_i)**.** *The probability of an $f_i \in \mathbb{F}_n^i$ is:*

$$P(f_i) = \frac{i!}{\prod_{j=1}^{i} f_{i,j}!} \binom{n-1}{i-1}^{-1}$$

Proposition (Probability of an f-sequence)**.**

$$P(f) = \prod_{i=2}^{n} P(f_i | f_{i-i}) \quad = \quad \frac{2^{\urcorner(f)}}{(n-1)!} \prod_{i=2}^{n} \ddot{f}_i$$

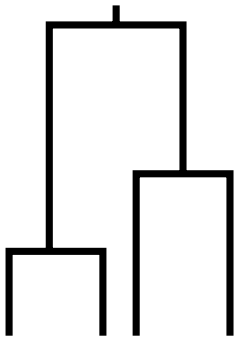
where,

- $\urcorner(f)$ *is the number of distinctly-sized lineage splits*
- $\ddot{f}_i$ *is the number of lineages at the beginning of the i-th epoch that subtend the same number of leaves as the lineage that was split then.*

c-sequence, $c \in \mathcal{C}_n \rightarrow$ c-shape, $\tilde{c} \in \widetilde{\mathcal{C}}_n \rightarrow$ f-sequence, $f \in \mathcal{F}_n$

$$n = 4$$

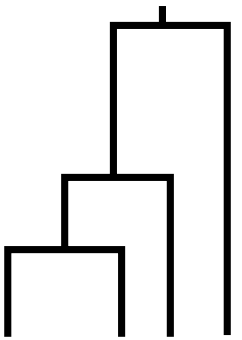
$\tilde{c}^{\wedge} = ((\cdot,_3\,\cdot),_1\,(\cdot,_2\,\cdot))$



$$\mathbf{f}^{\wedge} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$$

$\tilde{c}^{\curlywedge} = (((\cdot,_3\,\cdot),_2\,\cdot),_1\,\cdot)$

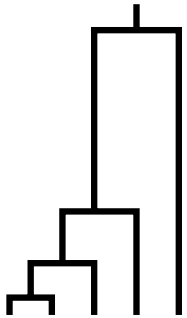
$$\mathbf{f}^{\curlywedge} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$$



c-sequence, $c \in \mathcal{C}_n \rightarrow$ c-shape, $\tilde{c} \in \widetilde{\mathcal{C}}_n \rightarrow$ f-sequence, $f \in \mathcal{F}_n$

$$n = 5$$

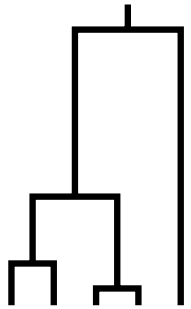
$\tilde{c}^{(\mathtt{a})} = ((((\cdot,_4\,\cdot),_3\,\cdot),_2\,\cdot),_1\,\cdot)$



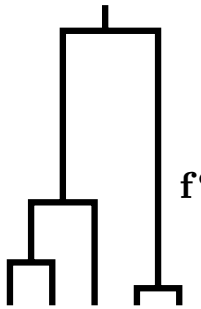
$$\mathbf{f}^{\mathtt{a}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$\tilde{c}^{(\mathtt{b})} = (((\cdot,_4\,\cdot),_2\,(\cdot,_3\,\cdot)),_1\,\cdot)$

$$\mathbf{f}^{\mathtt{b}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{pmatrix}$$

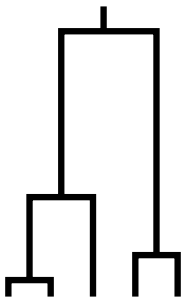


$\tilde{c}^{(\mathtt{c})} = ((\cdot,_4\,\cdot),_1\,((\cdot,_3\,\cdot),_2\,\cdot))$



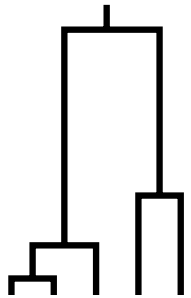
$\tilde{c}^{(\mathtt{d})} = (((\cdot,_4\,\cdot),_2\,\cdot),_1\,(\cdot,_3\,\cdot))$

$$\mathbf{f}^{\mathtt{c/d}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{pmatrix}$$



$\tilde{c}^{(\mathtt{e})} = (((\cdot,_4\,\cdot),_3\,\cdot),_1\,(\cdot,_2\,\cdot))$

$$\mathbf{f}^{\mathtt{e}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{pmatrix}$$



c-sequence, $c \in \mathcal{C}_n \rightarrow c$-shape, $\tilde{c} \in \widetilde{\mathcal{C}}_n \rightarrow f$-sequence, $f \in \mathcal{F}_n$

The number of c-sequences corresponding to the given f is

$$|F^{-1}(f)| = 2^{1-n}\, n!\, (n-1)!\, P(f) = n!\, 2^{\urcorner(f)+1-n} \prod_{i=2}^{n} \ddot{f}_i$$

Let $\beth(\tilde{c})$ be the number of cherries of a c-shape $\tilde{c} \in \widetilde{\mathcal{C}}$.

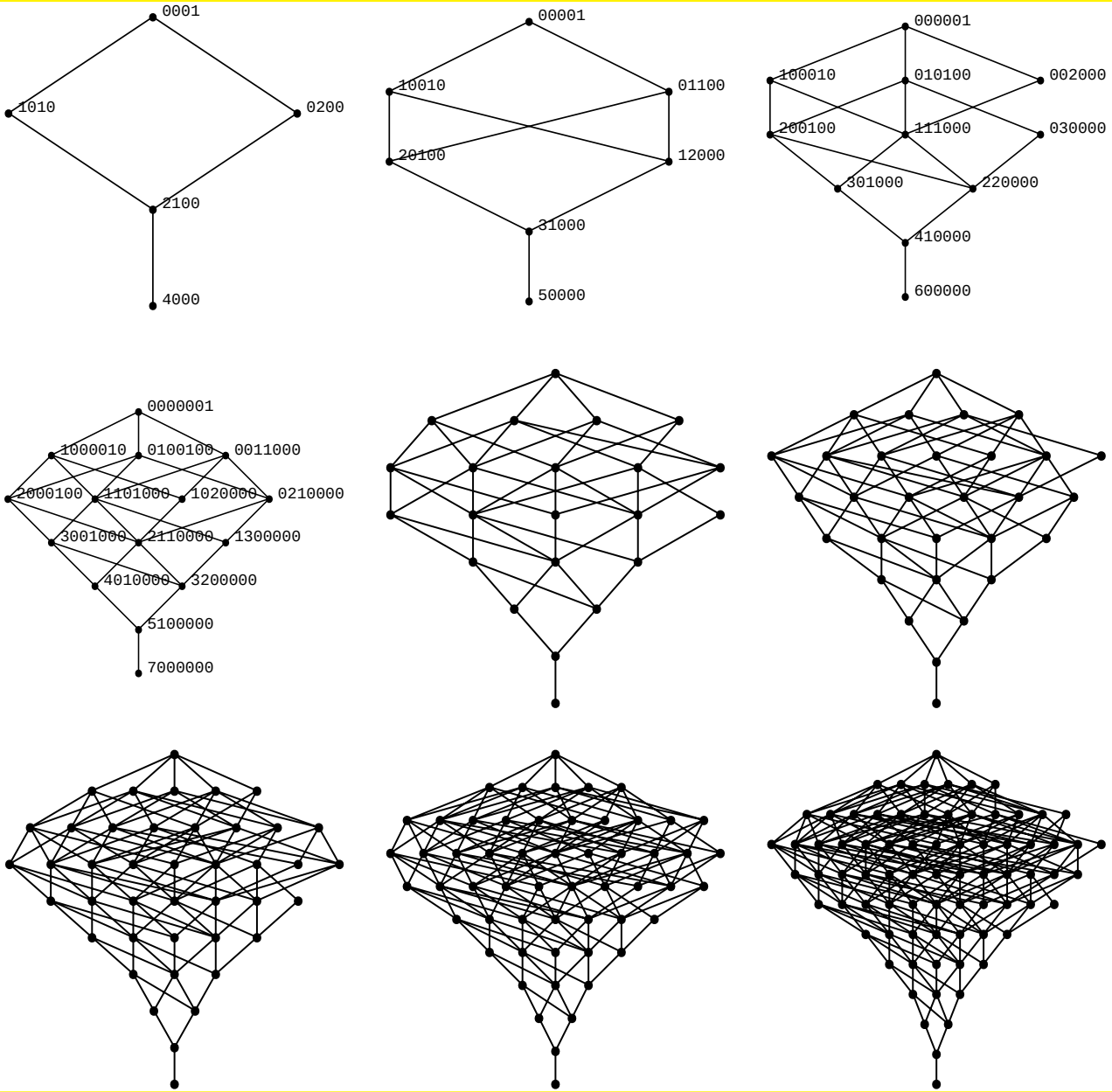
$$|\widetilde{C}^{-1}(\tilde{c})| = 2^{1-n}\, n!\, (n-1)!\, P(\tilde{c}) = n!\, 2^{-\beth(\tilde{c})} \quad (Tajima, 1983)$$

The number of c-shapes corresponding to the given f is

$$|\widetilde{C}(F^{-1}(f))| \;=\; 2^{-\beth(f)} \prod_{i=2}^{n} \ddot{f}_i,$$

$\beth(f) \equiv n - 1 - \urcorner(f) - \beth(f)$, the number of balanced splits that are not cherries.

Hasse Diagram of the Poset making $\mathcal{F}_n$ ($n = 4, \ldots, 12$)



Simulating f-sequences: for SFS, Shape Stats, ...

1: **input:**

   1. scaled mutation rate θ
   2. sample size n

2: **output:** a SFS sample x from the n-coalescent

3: generate an f-sequence under the unlabeled n-coalescent

4: draw $t \sim T = (T_2, T_3, \ldots, T_n)$, where T_i's are independently distributed as $\text{Exponential}\left(\binom{i}{2}\right)$

5: $l \leftarrow t^{\mathsf{T}} \cdot \mathbf{f}$ and $l_\bullet = \sum_{i=1}^{n-1}$

6: draw x from Poisson-Multinomial distribution $e^{-\theta l_\bullet}(\theta l_\bullet)^{\sum_{i=1}^{n-1} x_i} \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!$

7: **return:** x

Various tree shape statistics are further summaries of the f-sequence

$\tilde{s}$-sequence or Aldous shape statistic (Aldous, 2001)

$\widetilde{S}(f_n, f_{n-1}, \ldots, f_1) = \tilde{s} \equiv (\tilde{s}_n, \tilde{s}_{n-1}, \ldots, \tilde{s}_2) : \mathcal{F}_n \to \widetilde{\mathcal{S}}_n$:

$$\begin{aligned} \tilde{s}_i \equiv (\tilde{s}_{i,1}, \tilde{s}_{i,2}) &\equiv \left(\max\left(\|f\|_i\right), \min\left(\|f\|_i\right) 2^{-\mathbf{1}_{\{0\}}\left(\max\left(\|f\|_i\right) - \min\left(\|f\|_i\right)\right)}\right), \\ \|f\|_i &\equiv \{\, j | f_{i,j} - f_{i-1,j}| \in \mathbb{N} : j \in \{1, 2, \ldots, n\}\,\}. \end{aligned}$$

$$\mathfrak{Q}_n \equiv \{Q_{\mathtt{I}}(\tilde{s}) = q_{\mathtt{I}} \equiv \sum_{i=n}^{2} \tilde{s}_{i,1} \mathbf{1}_{\mathtt{I}}(\tilde{s}_{i,1}) : \widetilde{\mathcal{S}}_n \to \mathcal{Q}_{\mathtt{I}n},\ \mathtt{I} \in \mathbf{2}^{\{2,3,\ldots,n\}} \setminus \emptyset\}$$

$Q_{\{2,3,\ldots,n\}}(\tilde{s}) = q_{\{2,3,\ldots,n\}} = \sum_{i=n}^{2} \tilde{s}_{i,1}$ is the Sackin's index

$Q_{\{2\}}/2 = q_{\{2\}}/2$ is the number of cherries

$(n^2 - 3n + 2)^{-1} \sum_{i=n}^{2} (\tilde{s}_{i,1} - 2\tilde{s}_{i,d})$ is the Colless' index

Note: There are $2^{n-1} - 3$ others in the family $\mathfrak{Q}_n$

Likelihood of a Site Frequency Spectrum

Proposition (Likelihood of SFS)**.** *Let $a \in \mathcal{A}_n$ be a given coalescent tree, c be its c-sequence, $f = F(c)$ be its f-sequence, $t \equiv (t_2, t_3, t_n) \in (0, \infty)^{n-1}$ be its epoch times and let*

$$l \equiv (l_1, \ldots, l_{n-1}) = t^{\mathsf{T}} f = \left(\sum_{i=2}^{n} t_i f_{i,1}, \ldots \sum_{i=2}^{2} t_i f_{i,n-1} \right), \quad l_\bullet \equiv \sum_{i=2}^{n} l_i, \quad \bar{l}_i \equiv \frac{l_i}{l_\bullet}$$

be its lineage lengths subtending $1, 2, \ldots, n-1$ leaves, the total tree-size, and relative lineage lengths respectively.

Likelihood of a Site Frequency Spectrum

Proposition (Likelihood of SFS)**.** *Let $a \in \mathcal{A}_n$ be a given coalescent tree, c be its c-sequence, $f = F(c)$ be its f-sequence, $t \equiv (t_2, t_3, t_n) \in (0, \infty)^{n-1}$ be its epoch times and let*

$$l \equiv (l_1, \ldots, l_{n-1}) = t^{\mathsf{T}} f = \left(\sum_{i=2}^{n} t_i f_{i,1}, \ldots \sum_{i=2}^{2} t_i f_{i,n-1} \right), \quad l_\bullet \equiv \sum_{i=2}^{n} l_i, \quad \bar{l}_i \equiv \frac{l_i}{l_\bullet}$$

be its lineage lengths subtending $1, 2, \ldots, n-1$ leaves, the total tree-size, and relative lineage lengths respectively.

$$P(x|\phi, a) \quad = \quad P(x|\phi, l = t^{\mathsf{T}} f) = e^{-\theta l_\bullet} (\theta l_\bullet)^S \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!$$

Likelihood of a Site Frequency Spectrum

Proposition (Likelihood of SFS). *Let $a \in \mathcal{A}_n$ be a given coalescent tree, c be its c-sequence, $f = F(c)$ be its f-sequence, $t \equiv (t_2, t_3, t_n) \in (0,\infty)^{n-1}$ be its epoch times and let*

$$l \equiv (l_1, \ldots, l_{n-1}) = t^{\mathsf{T}} f = \left(\sum_{i=2}^{n} t_i f_{i,1}, \ldots \sum_{i=2}^{2} t_i f_{i,n-1} \right), \quad l_\bullet \equiv \sum_{i=2}^{n} l_i, \quad \bar{l}_i \equiv \frac{l_i}{l_\bullet}$$

be its lineage lengths subtending $1, 2, \ldots, n-1$ leaves, the total tree-size, and relative lineage lengths respectively.

$$P(x|\phi, a) = P(x|\phi, l = t^{\mathsf{T}} f) = e^{-\theta l_\bullet} (\theta l_\bullet)^S \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!$$

$$P(x|\phi) = \frac{1}{\prod_{i=1}^{n-1} x_i!} \sum_{f \in F_n^c(x^\circledast)} P(f) \left(\int_{t \in (0,\infty)^{n-1}} \left(e^{-\theta l_\bullet} (\theta l_\bullet)^S \prod_{i=1}^{n-1} \bar{l}_i^{x_i} \right) P(t|\phi) \right)$$

$$\text{where,} \quad F_n(x^\circledast) \equiv \bigcup_{\{h : x_h^\circledast = 1\}} \{ f \in \mathcal{F}_n : \sum_{i=1}^{n} f_{i,h} = 0 \}$$

$$X^\circledast(x) = x^\circledast \equiv (x_1^\circledast, \ldots, x_{n-1}^\circledast) \equiv (\, \mathbf{1}_{\mathbb{N}}(x_1), \ldots, \mathbf{1}_{\mathbb{N}}(x_{n-1}) \,) \in \{0,1\}^{n-1}$$

An Importance Sampler over $F_n^c(x^{\circledast})$

Proposition (A Proposal over $F_n^c(x^{\circledast})$)**.** *For a given $x \in \mathcal{X}_n^m$, consider the following discrete time Markov chain on the augmented state space $\mathbb{F}_n \times \{0,1\}^{n-1} \ni (f_h, z_h)$:*

$$P^*((f_h, z_h)|(f_g, z_g)) = \begin{cases} P(f_h|f_g)/\Sigma(f_g, z_g) & \text{: if} \quad (f_h, z_h) \prec_{f,z} (f_g, z_g), \\ 0 & \text{: otherwise} \end{cases}$$

where,

$$\Sigma(f_g, z_g) = \sum_{(j,k) \in H(f_g, z_g)} P(f_g - e_{j+k} + e_j + e_k | f_g),$$

$$H(f_g, z_g) = \{(j,k) : f_{g,j+k} > 0,\ 1 \le j \le \max\{\min\{\hat{g}, j+k-1\}, \lceil \frac{j+k}{2} \rceil\} \le k \le j+k-1\},$$

$$\hat{g} = \max\{i : z_{g,i} = 1\},$$

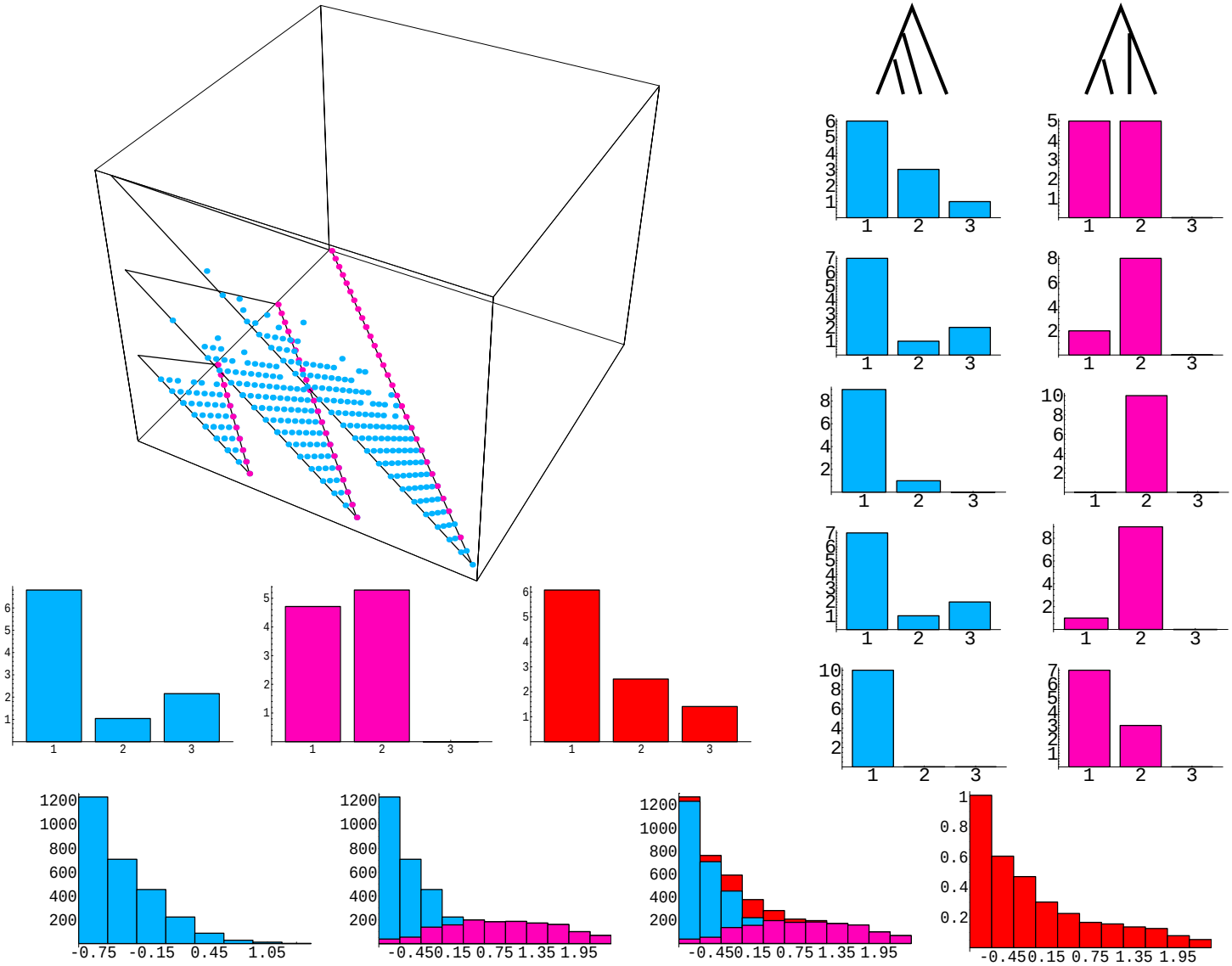
$$(f_h, z_h) \prec_{f,z} (f_g, z_g) \Leftrightarrow f_h = f_g + e_j + e_k - e_{j+k},\ z_h = z_g - \mathbf{1}_{\{1\}}(z_{g,j})\, e_j - \mathbf{1}_{\{1\}}(z_{g,k})\, e_k$$

where, the initial state is $(f_1, X^{\circledast}(x)) = ((0,0,\ldots,1), x^{\circledast})$ and the final absorbing state is $(f_n, (0,0,\ldots,0)) = ((n,0,\ldots,0),(0,0,\ldots,0))$.

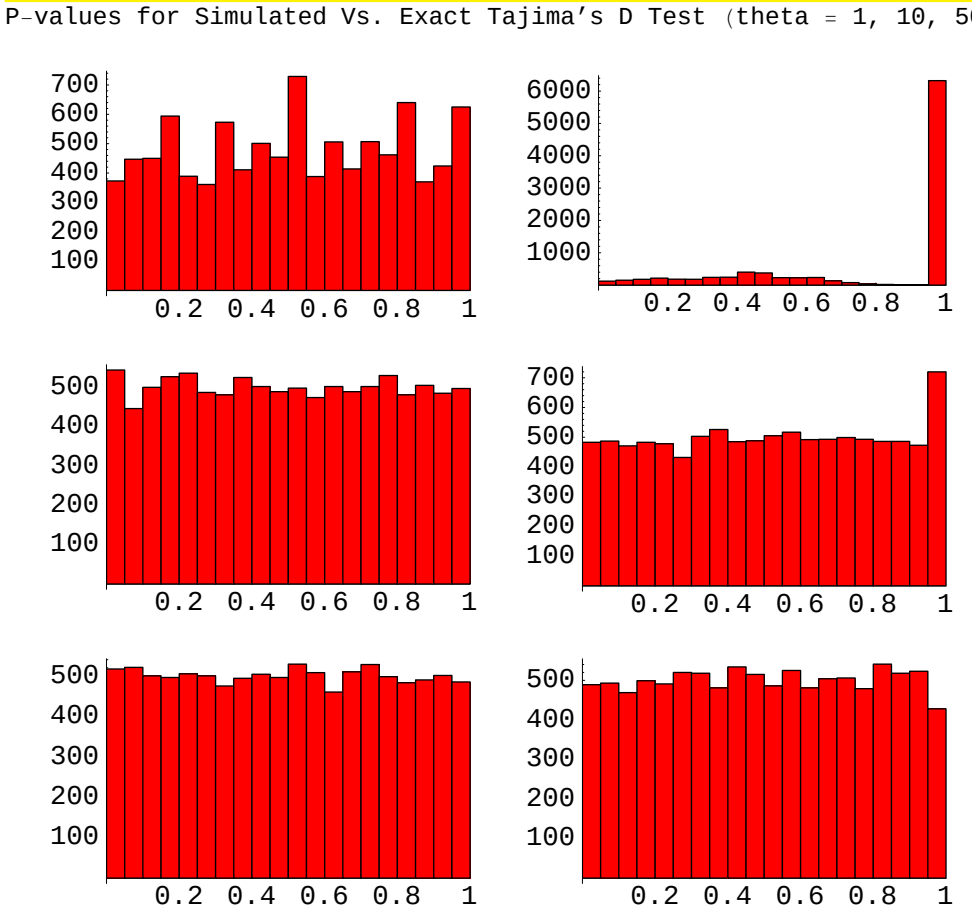
Maximum *Aposteriori* Estimates of θ and ν by $\sum$ over $f \in F_n^c(x^{\circledast})$

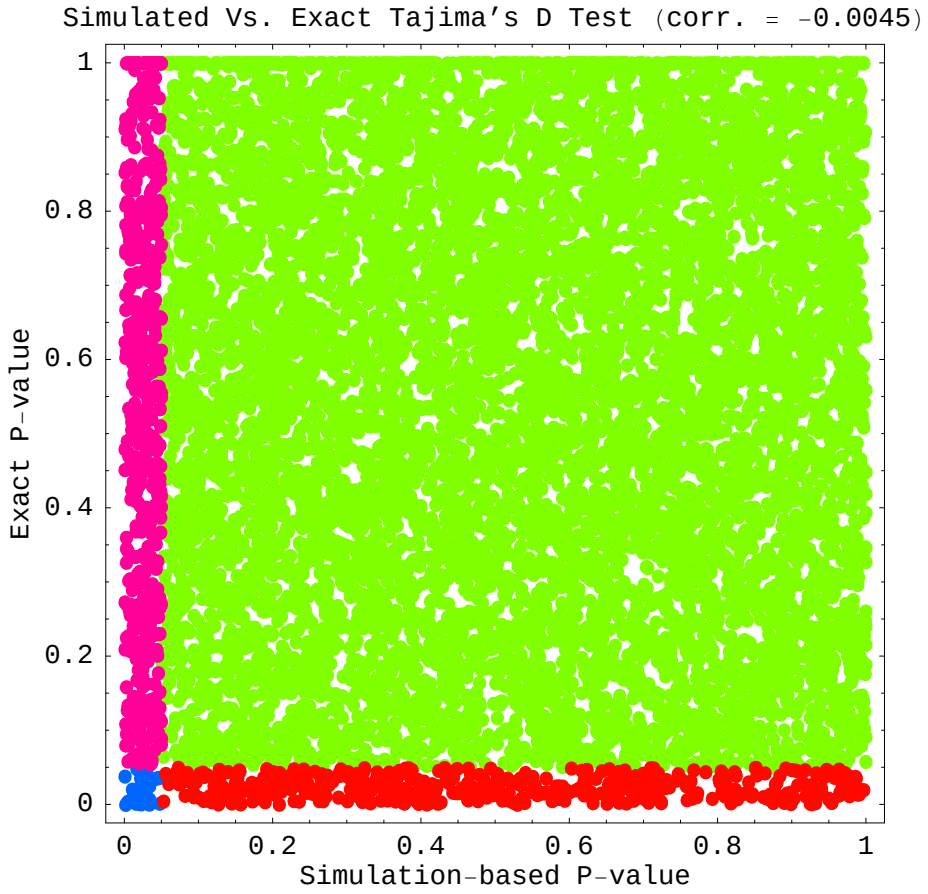
n	$\widehat{\nu}$			$\widehat{\theta}$			$(\widehat{\theta}, \widehat{\nu})$	
	$\sqrt{se}$	bs	$C_{99\%}$	$\sqrt{se}$	bs	$C_{99\%}$	$C_{99\%}$	$Qrt(\breve{\mathcal{K}})$
4	46	30	42	43	30	53	98	$\{0.061, 0.079, 0.13\}$
5	32	19	42	31	22	63	96	$\{0.074, 0.098, 0.16\}$
6	31	18	41	35	23	69	93	$\{0.082, 0.11, 0.17\}$
7	34	19	48	32	20	68	87	$\{0.090, 0.12, 0.21\}$
8	26	12	66	21	11	72	92	$\{0.098, 0.14, 0.26\}$
9	27	12	65	18	10	70	93	$\{0.097, 0.14, 0.21\}$
10	23	11	64	17	10	66	95	$\{0.091, 0.14, 0.30\}$

Topological Unfolding of SFS and Tajima's D when $n = 4$



Simulated Vs. Gen. Fisher's Exact Test with Tajima's D

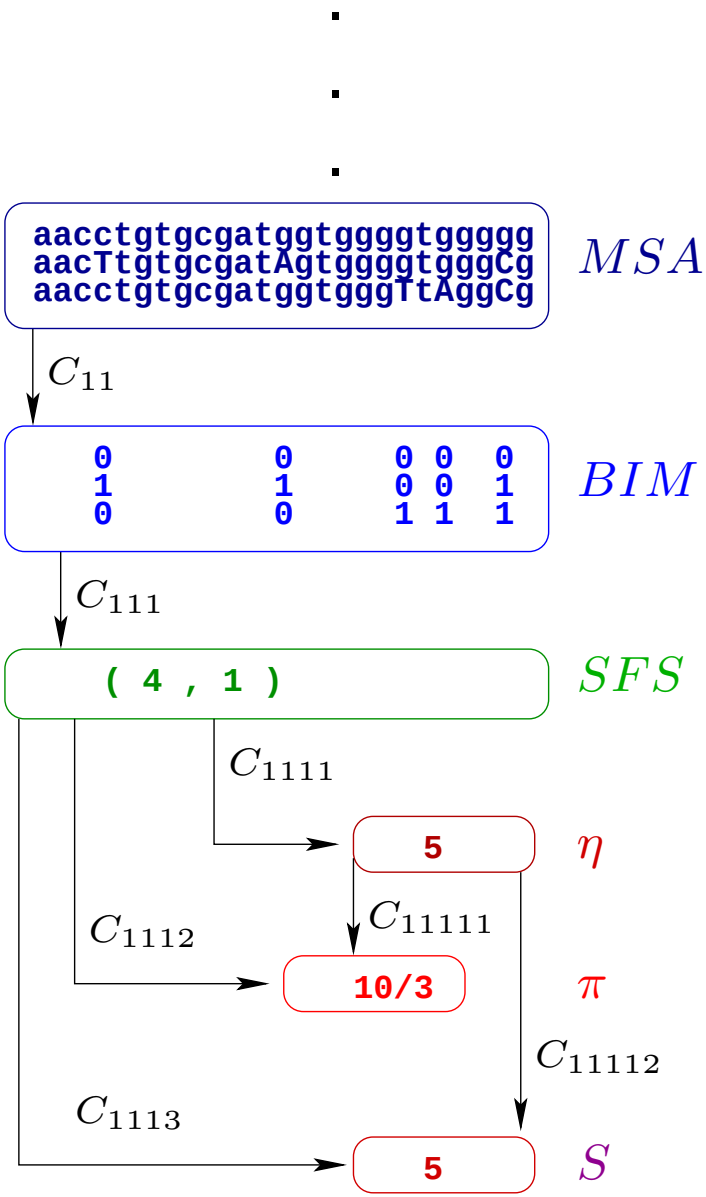




Left panel: Distribution of p-values from the simulated test (left) and the generalized Fisher's exact test (right) for three values of $\theta = \{1, 10, 50\}$ per 1000 bp with $n = 30$.

Right panel: The almost zero correlation of p-values between the two tests.

Summary



- Limits on Inference from Finest Empirical Resolutions
- Inference from Coarser Site Frequency Spectrum is Possible via a Collapsed Kingman's n-coalescent Markov chain
- Algebraic Geometry is useful to infer from classical summaries of SFS.
- MSEs are smaller – the exponential growth model
- Helps speed-up intensive SIS methods (Particle filtering on Experiment Graph)
- Topological unfolding of SFS and D $\Rightarrow$ Tree-less Genome Scans are essentially meaningless
- A Decision-theoretic formalism – partially-ordered coalescent experiments graph
- Possible to generalize
- Saves electricity and slows down global warming!

- NSF/NIGMS grant DMS-02-01037 to Durrett, Aquadro, and Nielsen and
- Research Fellow of the Royal Commission for the Exhibition of 1851.